INTRODUCTION TO DATA SCIENCE

ASSIGNMENT PART 1

Group 079

Fahimeh Fereydounian

Aroma Agarwal

Joost-Henning Groot Bramel

QUESTION: 1

Data Exploration

1a)

920 rows × 16 columns

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
915	916	54	Female	VA Long Beach	asymptomatic	127.0	333.0	True	st-t abnormality	154.0	False	0.0	NaN	NaN
916	917	62	Male	VA Long Beach	typical angina	NaN	139.0	False	st-t abnormality	NaN	NaN	NaN	NaN	NaN
917	918	55	Male	VA Long Beach	asymptomatic	122.0	223.0	True	st-t abnormality	100.0	False	0.0	NaN	NaN
918	919	58	Male	VA Long Beach	asymptomatic	NaN	385.0	True	lv hypertrophy	NaN	NaN	NaN	NaN	NaN
919	920	62	Male	VA Long Beach	atypical angina	120.0	254.0	False	lv hypertrophy	93.0	True	0.0	NaN	NaN

920 rows × 16 columns

1b)

Some of the columns are not numerical and these computations do not apply to categorical features.

	id	age	trestbps	chol	thalch	oldpeak	ca	num
count	920.000000	920.000000	861.000000	890.000000	865.000000	858.000000	309.000000	920.000000
mean	460.500000	53.510870	132.132404	199.130337	137.545665	0.878788	0.676375	0.995652
std	265.725422	9.424685	19.066070	110.780810	25.926276	1.091226	0.935653	1.142693
min	1.000000	28.000000	0.000000	0.000000	60.000000	-2.600000	0.000000	0.000000
25%	230.750000	47.000000	120.000000	175.000000	120.000000	0.000000	0.000000	0.000000
50%	460.500000	54.000000	130.000000	223.000000	140.000000	0.500000	0.000000	1.000000
75%	690.250000	60.000000	140.000000	268.000000	157.000000	1.500000	1.000000	2.000000
max	920.000000	77.000000	200.000000	603.000000	202.000000	6.200000	3.000000	4.000000

```
Total number of NaN values: 1759
```

1c)

The values in the num column are mostly close to 0.996, with a median of 1, slightly higher than the mean, indicating a left-skewed distribution likely influenced by variability or outliers. Therefore, relying on the mean alone is insufficient, and using the median alongside the mean provides a more accurate representation of the distribution.

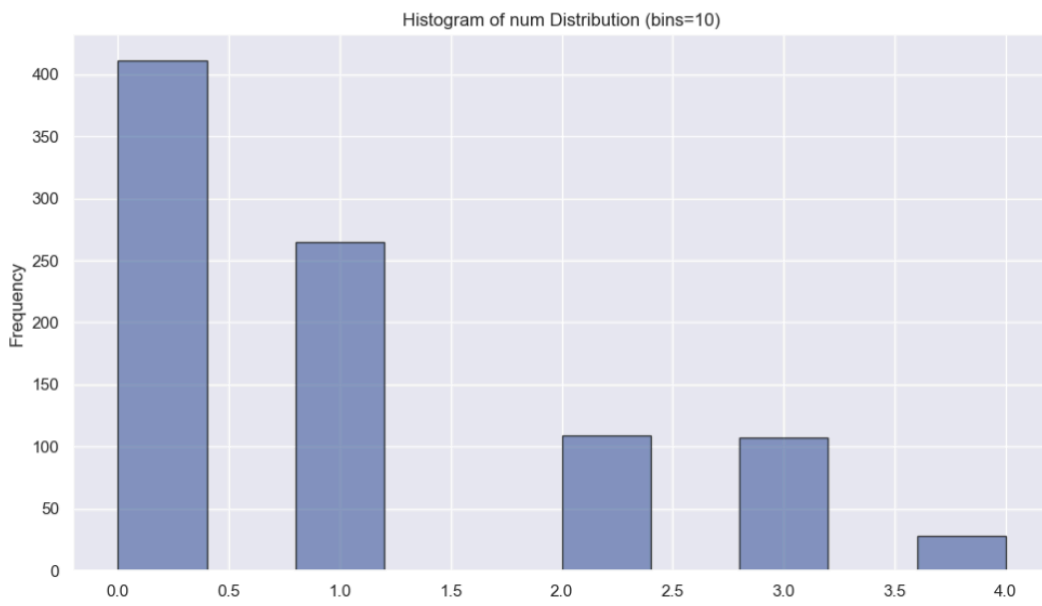
1d)

```
Count of each value in 'sex':
sex
Male      726
Female    194
```

To balance out their representation, we can use **undersampling or oversampling.**

1e)

Mode: 0

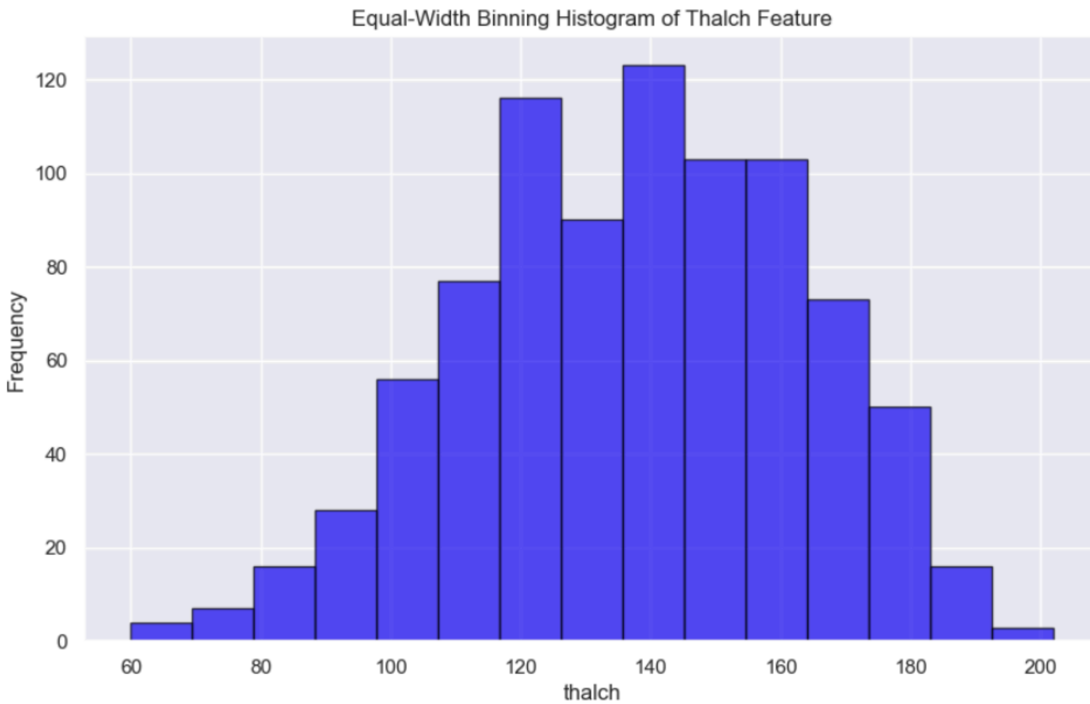


1f)

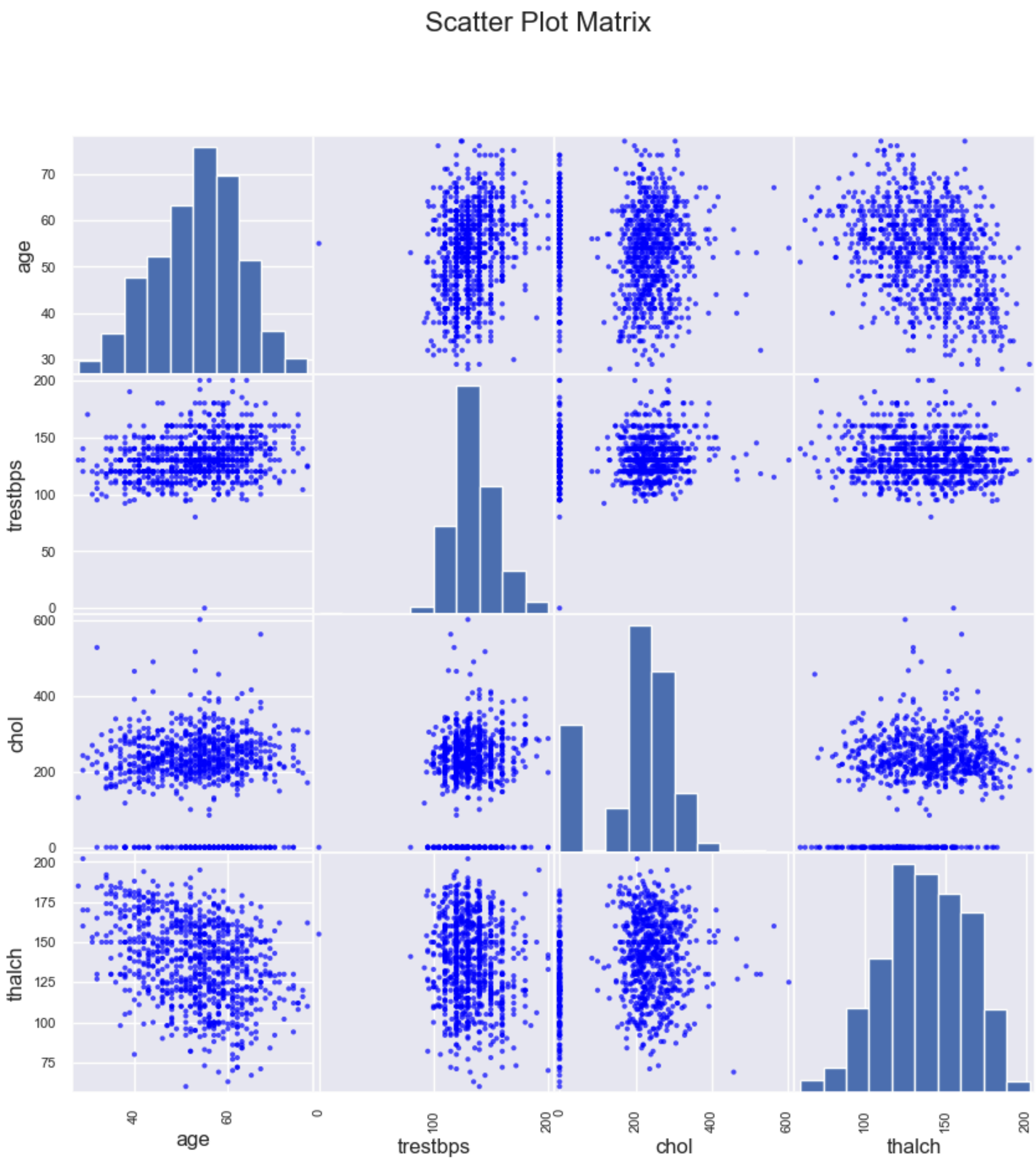
Number of bins: 15

I started using different numbers for bins, and I realized numbers larger than 20 are not adding detail to the data. I searched for different methods and found that Scott's Rule is a good method to determine the number of bins for the distribution of thalch values.

The result is a **multimoal distribution histogram** and it has two peaks.


Equal-Width Binning Histogram of Thalch Feature
Frequency
thalch
0
20
40
60
80
100
120
60
80
100
120
140
160
180
200

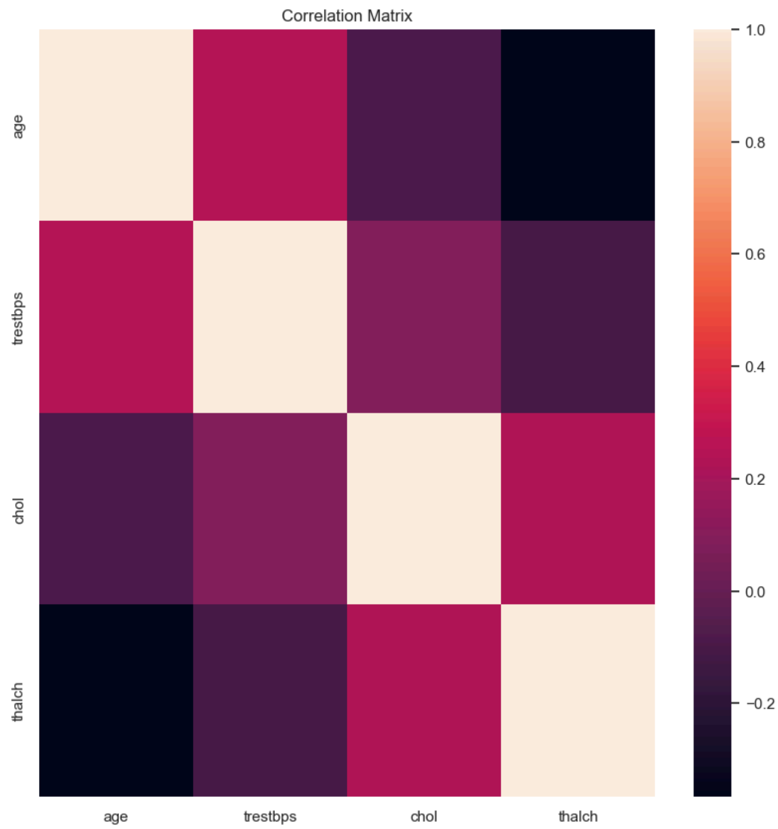
1g)



1h)

The strongest absolute correlation is the **correlation between age and thalch, which equals -0.366**, which is a (weak) negative correlation. It means that as age increases, the maximum heart rate decreases. (r < 0 shows negative relationship between variables.)

	age	trestbps	chol	thalch
age	1.000000	0.244253	-0.086234	-0.365778
trestbps	0.244253	1.000000	0.092853	-0.104899
chol	-0.086234	0.092853	1.000000	0.236121
thalch	-0.365778	-0.104899	0.236121	1.000000



1i)

- **We can compare the boxplot but cannot draw a precise conclusion by just looking at them.**
  Female: The median is slightly above the median in the male boxplot. Similar to 1.c, this boxplot is also skewed towards left, which means mean (or average) < median. There are some outliers and some of them are below the lower fence.
  Male: It is almost symmetrical, which means median is close to mean. However, there are more outliers above the upper whisker, which increases the average. So here average > median.
  But we don't know exactly if the mean of female, which is lower than the median of female, will also be lower than the mean in male. **I estimated the quartiles and whiskers using rounded numbers, and it appears that the male mean is marginally higher than the female mean. However, this could not be inferred from the first look at box plots.**
- No.
- Cannot be answered. The number of outliers below the lower whisker in the female group is unclear (apparently there are more than one drawn on the same spot).
- Yes.

1j)

Total number of NaN values: 1759

fbs feature has the most NaNs: 692

Lower bound: minimum number of rows that must contain NaNs if they are densely packed. -> 920 - 46 = 874

Upper bound: maximum number of rows that must contain NaNs if they are spread out in rows. -> 920 (Adding up the NaNs in total will exceed the total number of rows, so the upper bound is 920.)

```
Features' NaNs

id            0
age           0
sex           0
dataset       0
cp            0
trestbps      1
chol        172
fbs         692
restecg       0
thalch        0
exang       528
oldpeak     370
slope         0
ca          181
thal          0
num         411
```

```
Number of NaNs per entry

4    246
2    215
1    200
3    162
5     51
0     46
Name: count, dtype: int64

Total number of rows: 920
```

1k)

The number of rows changes from 920 to 299.

QUESTION: 2

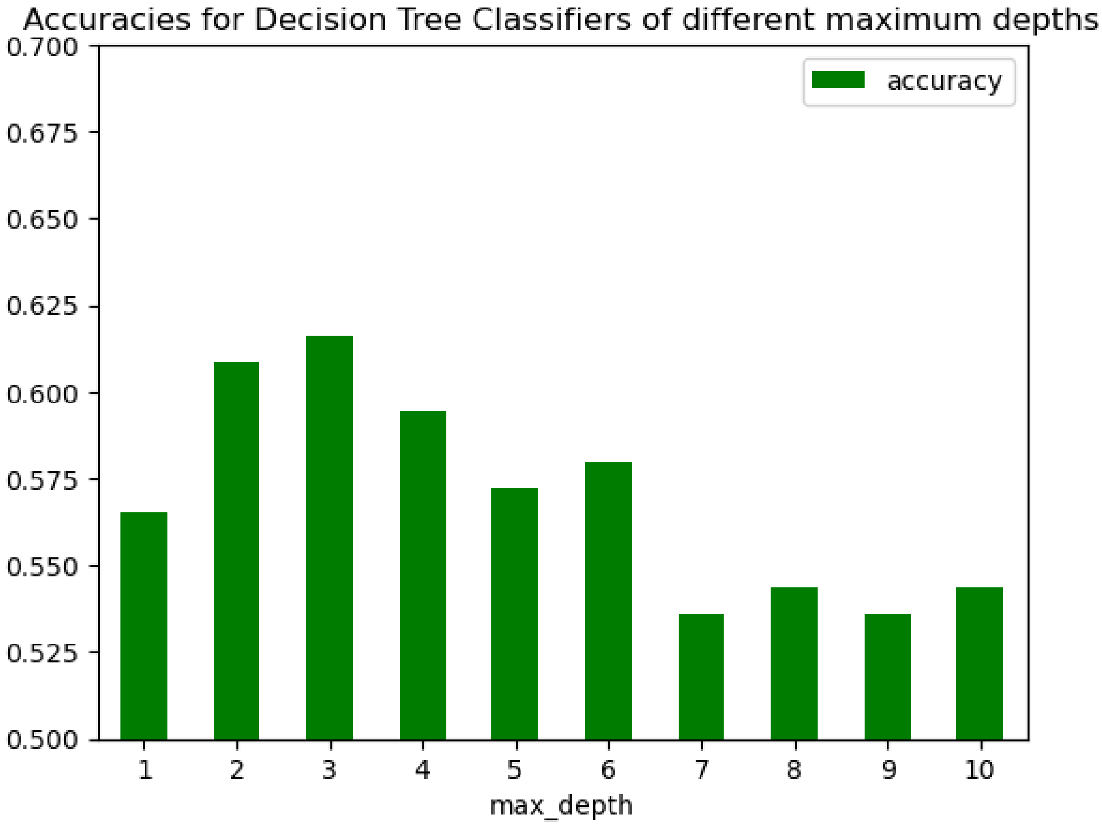
Decision Trees

2a) Baseline

For the baseline we used the sklearn DummyClassifier, which ignores the input data and randomly predicts given one of several strategies. With an accuracy of 0.29 the DummyClassifier is basically a little better then guessing.

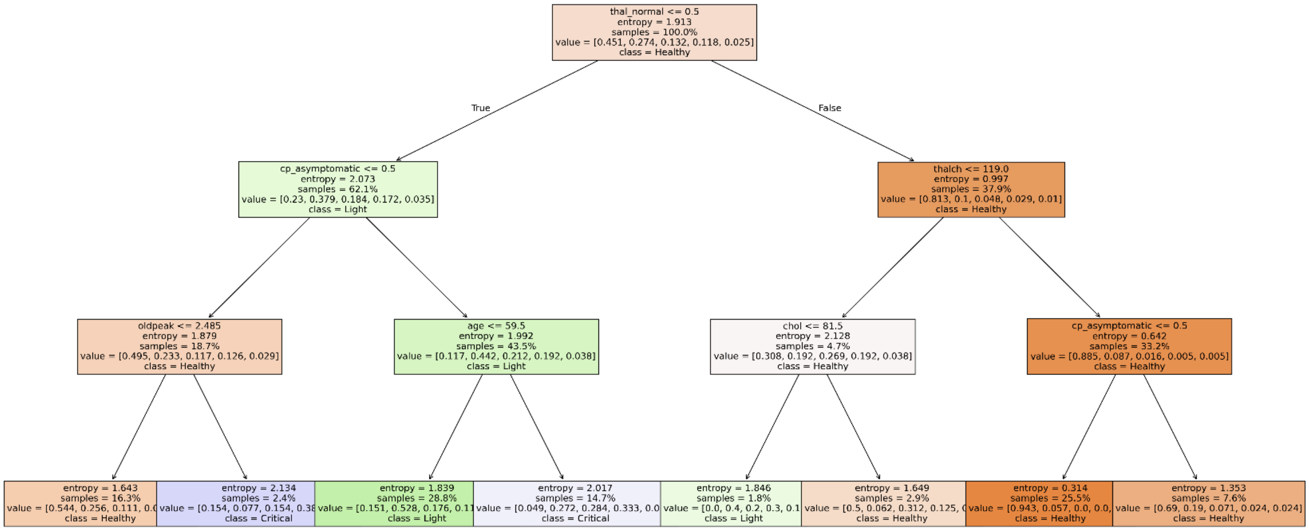
```
Baseline Classifier Accuracy: 0.2971014492753623
Classification Report:
              precision    recall  f1-score   support

           0       0.46      0.45      0.46        71
           1       0.16      0.12      0.14        42
           2       0.00      0.00      0.00         8
           3       0.23      0.21      0.22        14
           4       0.17      0.33      0.22         3

    accuracy                           0.30       138
   macro avg       0.20      0.22      0.21       138
weighted avg       0.31      0.30      0.30       138
```

2b)



In the plot above we plotted the accuracy of the decision tree classifier in dependence of its maximum depth. We would choose a max_depth of 3 with the highest accuracy of 0.615942.

2c)



The first decision criterion is, if the value in the thal_normal column is less or equal 0.5. The thalch states the maximum heart rate achieved in a stress test and thal_normal states how far the rate deviates from healthy individuals.

QUESTION: 3

Regression

3a)

Weights:

[[-0.03595986 0.03282182]

 [-0.01361616 -0.01513819]

 [0.04444063 -0.01085738]

 [0.05601142 -0.02683136]

 [0.09108474 -0.01093397]]

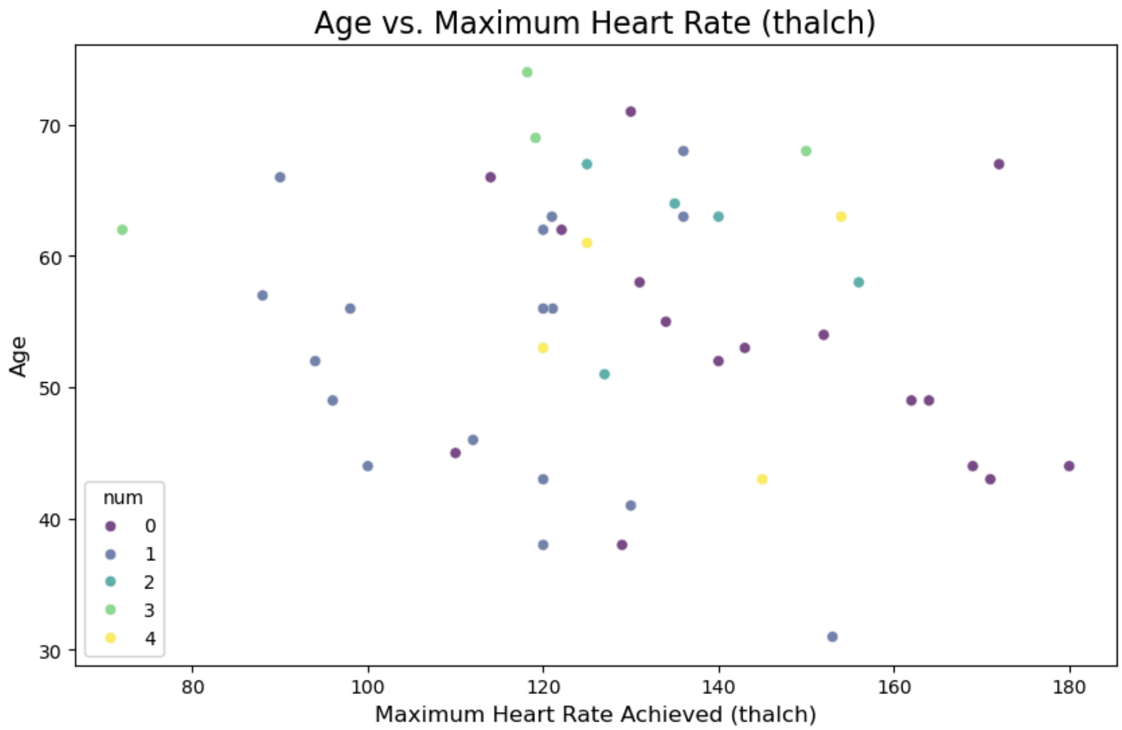
Intercept:

 [-2.76868524 1.8326611 -3.02219145 -1.70964226 -7.46803626]

3b)

Mean Squared Error 2.260869565217391
Accuracy Score 0.5217391304347826

3c)

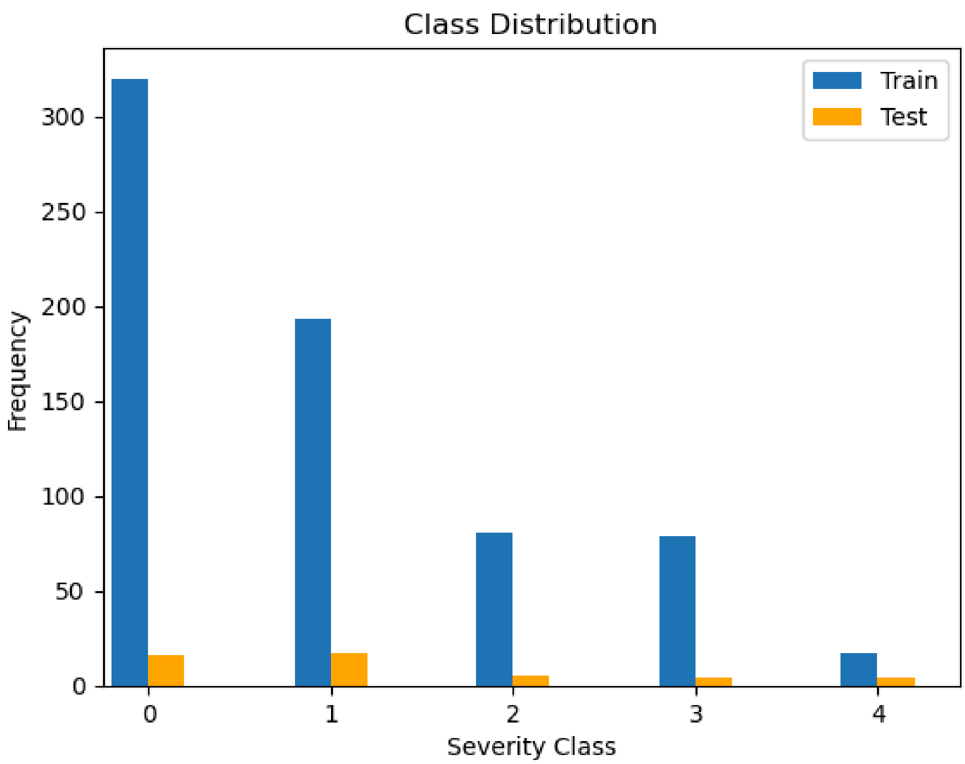


The performance of the Logistic Regression Model is bad, because the groups have large overlaps and there are no clear borders visible. We could try to find another descriptive feature in our data, that might add another dimension to our classification problem. One could also try to balance the number of class occurrences better or only test for one class at a time.

QUESTION: 4

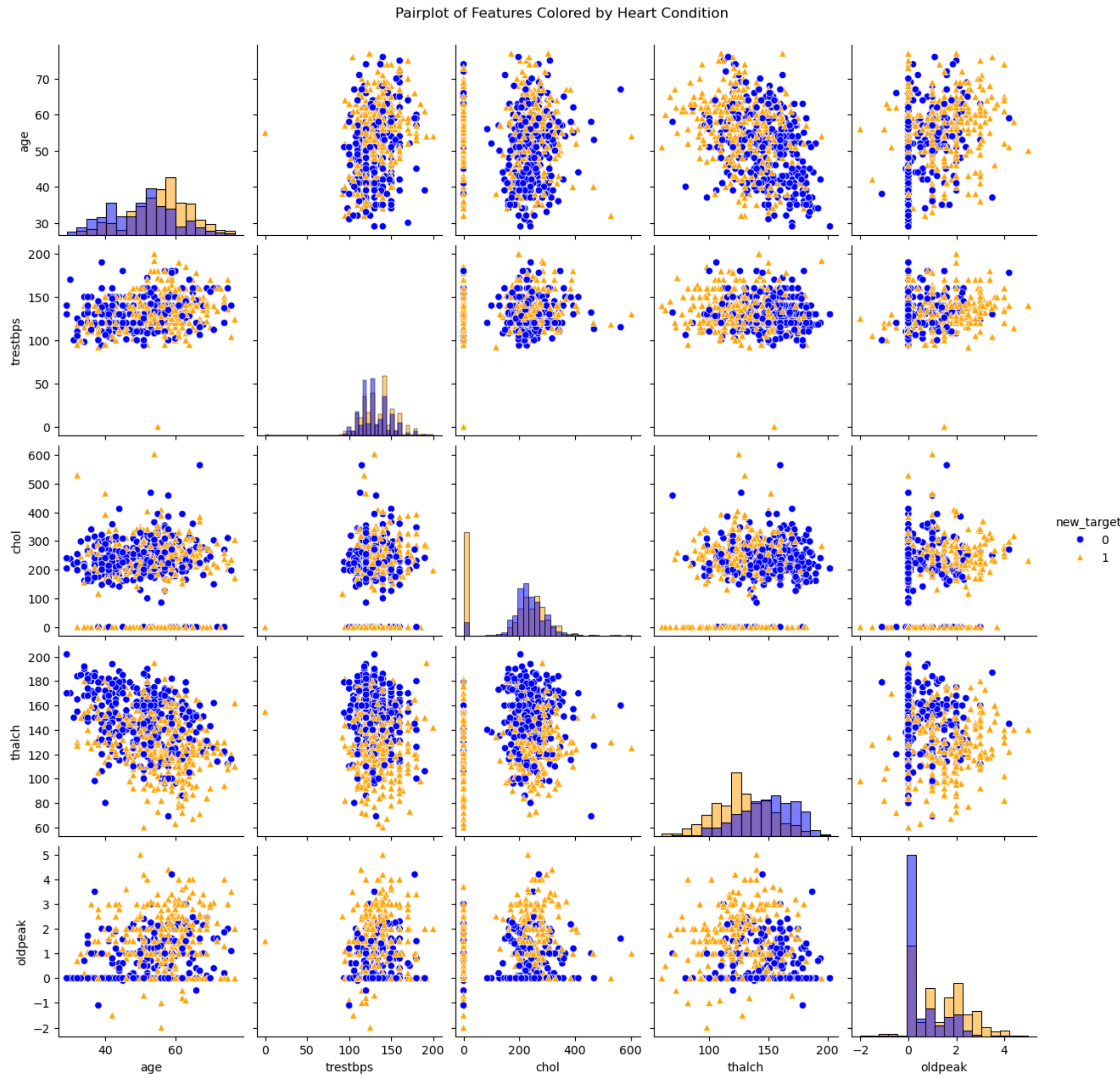
Support Vector Machines and Neural Networks

4a)

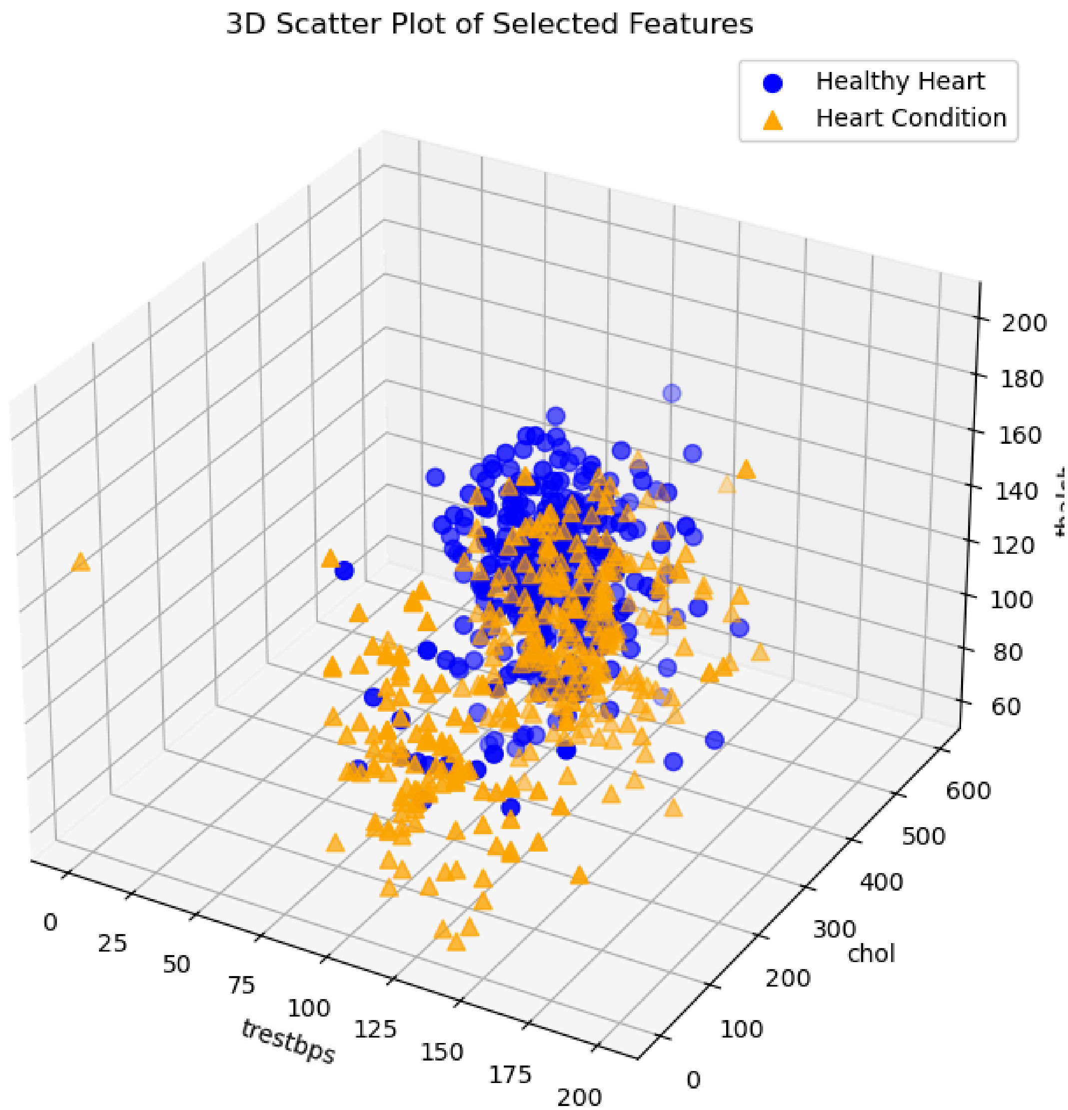


For Training Data:
Class 0 = 320 instances, Class 1 = 193 instances. The instances keep decreasing with severity, Class 4 has the least amount of instances (17).
For Test Data:
Class 0(16) and Class 1(17) have almost equal instances. Classes 2, 3, 4 have very few instances.
Class distribution for both datasets is similar but they differ in scale.

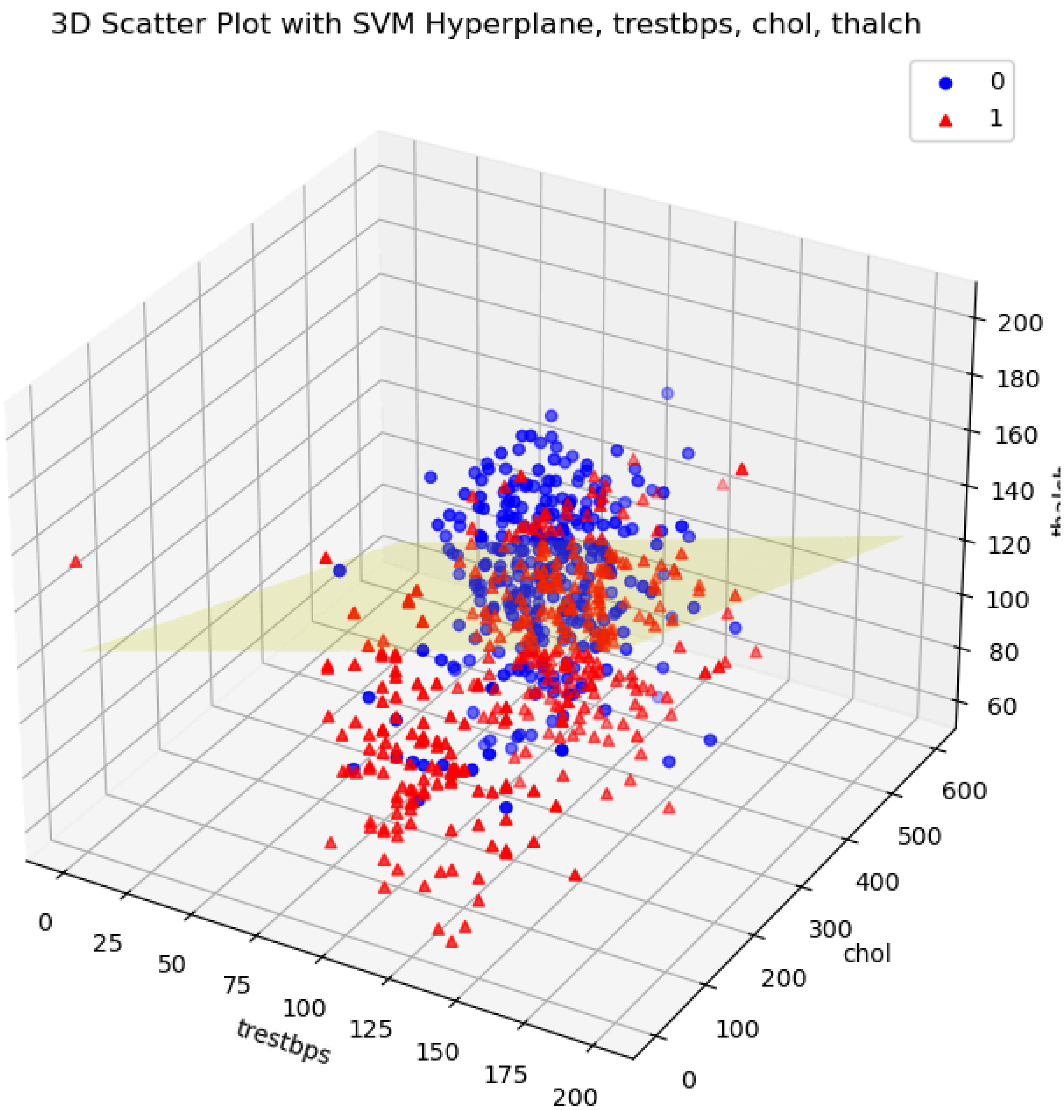
4b)



Pairplot of Features Colored by Heart Condition
age
trestbps
chol
thalch
oldpeak
new_target
0
1



The selected features are *trestbps*, *chol* and *thalch*. On observing the pair plot and the the 3D Plot, these three features seemed to give the best separation between the classe, so we decided to select these three features for further tasks.



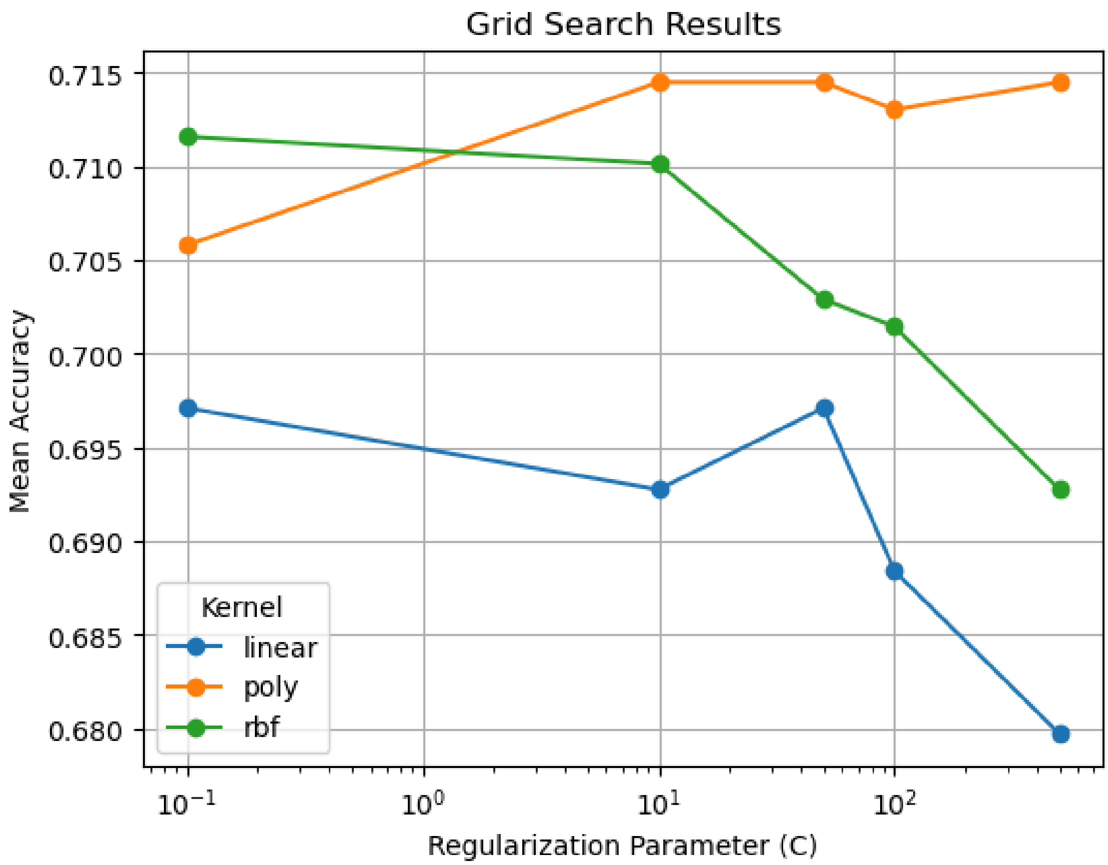
(3D plot with hyperplane using the given function)

4c)

Model Accuracy: 0.739
Model Precision: 0.781
Given the size of the training dataset, the model seems to be working quite well. With larger samples of training and test datasets, the accuracy and precision of the model could be improved

4d)

The best-performing configuration is the Polynomial Kernel function with the Regularization parameter of 10.



The linear kernel performs significantly worse than the other two kernels for all the regularisation parameter.

4e)

Linear kernel, C=50 : Accuracy = 0.74, Precision = 0.76
RBF Kernel, C=0.1: Accuracy = 0.65, Precision = 0.65
Polynomial Kernel, C=10: Accuracy = 0.74, Precision = 0.76
Although, Linear model performed the worse in Grid Search, it performs the best alongside Polynomial Kernel. Also, on the basis of the Grid Search Results, the Accuracy of the model decreases with the increase in Regularization Parameter for Linear and RBF Kernel, but increases for Polynomial Model.

4f)

Accuracy: 0.83
Precision: 0.87
Confusion Matrix:
[[12   4]
 [4   26]]

While considering all the features, the accuracy of the model increases.
Considering the ratio of true positive to false positive and true negative to false negative, the model favours the no heart condition class.

The percentage of patients classified as healthy but have heart conditions is 13.33%.

4g)

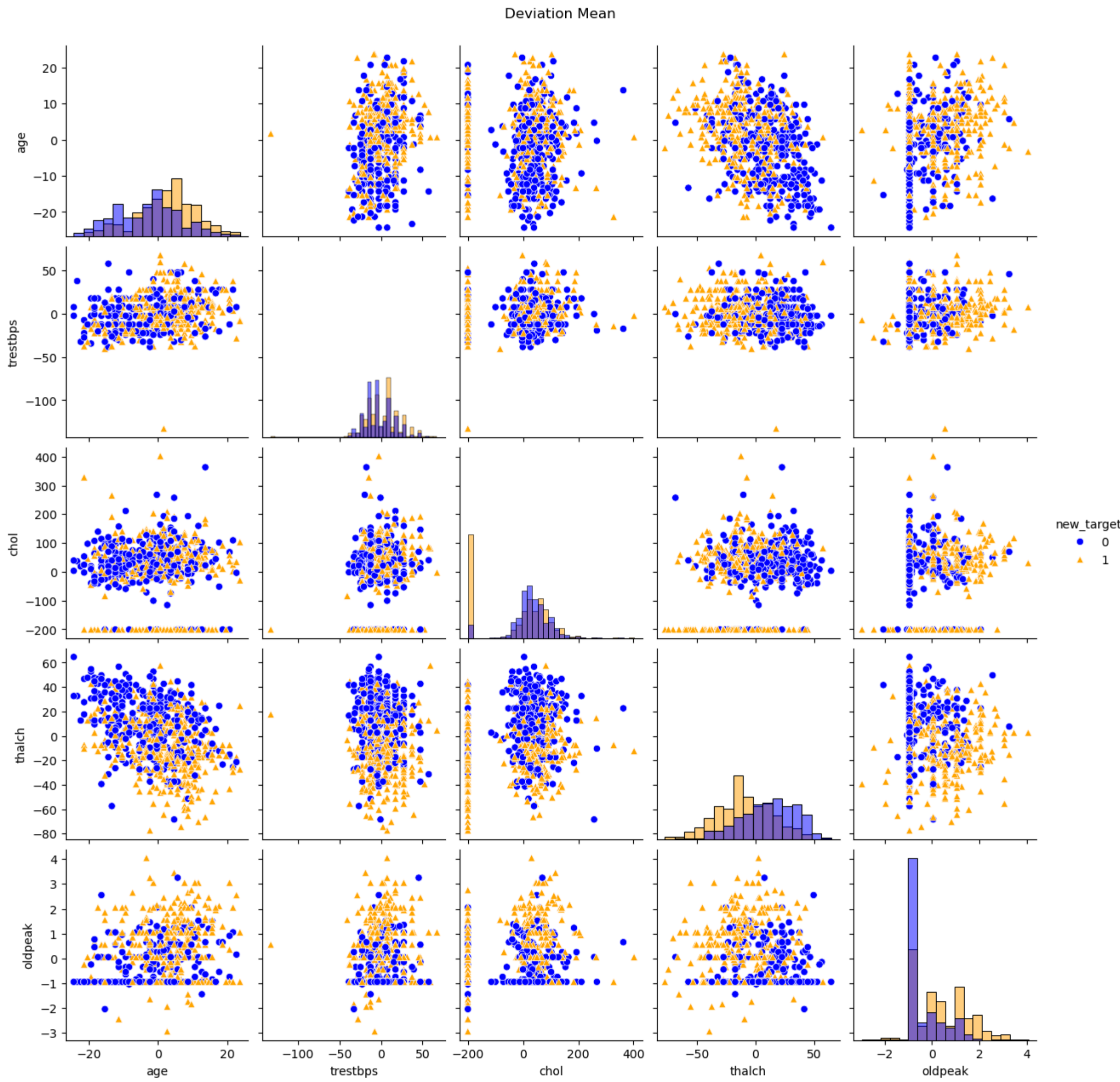
Accuracy: 0.35
[[16 0 0 0 0]
 [17 0 0 0 0]
 [5 0 0 0 0]
 [4 0 0 0 0]
 [4 0 0 0 0]]

All the patients are categorised as having no health condition. This could be because the sample training data set is not well distributed and is pretty small. Moreover, the test data set is also pretty skewed, which is why we get a pretty low accuracy.

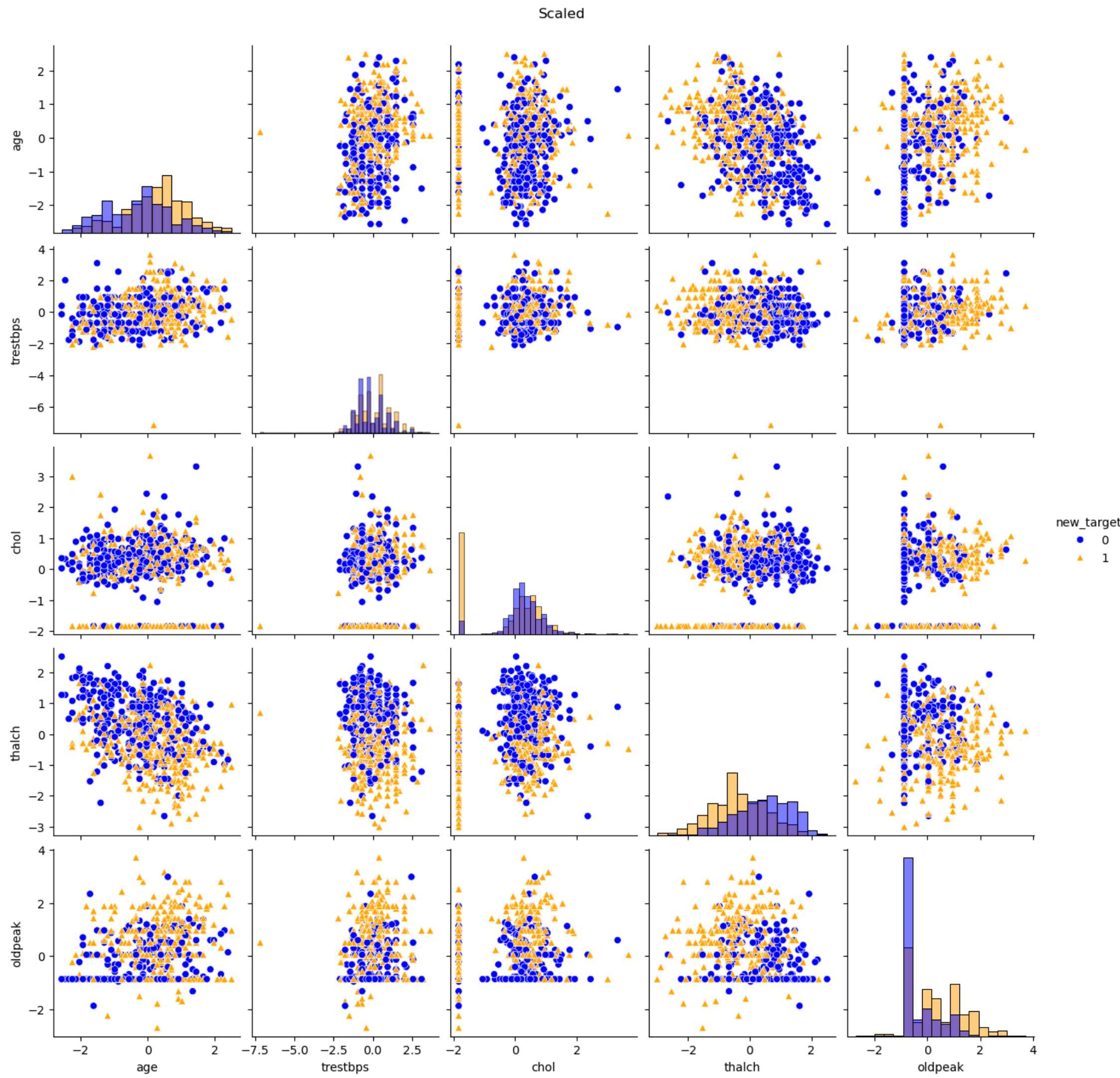
4h)

Feature engineering helps in transforming data to improve a model's ability to identify patterns and thus improving its performance.

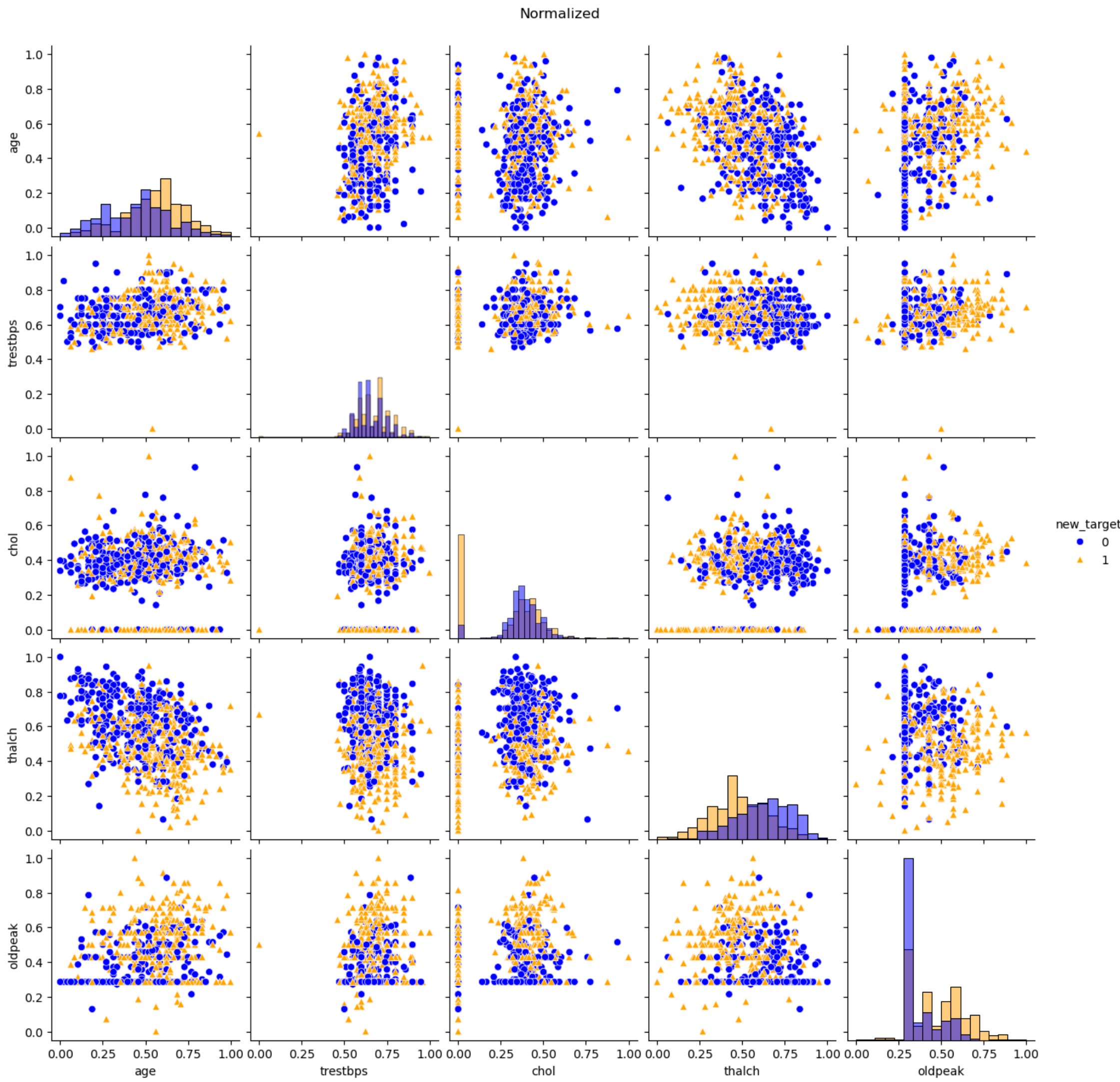
Paiplot with data preprocessed using the deviation mean



Paiplot with data preprocessed by scaling



Paiplot with data preprocessed by normalisation



Scaling standardizes the magnitude, and normalisation also eliminates magnitude difference. Normalisation transforms features in the range of 0-1. Scaling and deviation mean both adjust the data around zero, and deviation mean just doesn't adjust the scales.

Scaling ensures that no feature dominates due to large magnitudes, so it helps SVMs. Normalisation and scaling also aid Neural Networks which use gradient-based optimisation

4i)

Accuracy: 0.50
[[14 2 0 0 0]
 [8 9 0 0 0]
 [2 3 0 0 0]
 [0 4 0 0 0]
 [0 4 0 0 0]]

Hidden Layers = 3 with each having 9 neurons

Accuracy was peaking at 8 or 9 neurons and even after repeated trials, it could not detect classes apart from 0 and 1. On increasing hidden layers to more than 3, in most cases, it is not even able to detect class 1. So we decided to settle for these adjustments.

The accuracy of the model does increase from 0.35 earlier to 0.5.

4j)

Accuracy: 0.54
[[14 2 0 0 0]
 [6 11 0 0 0]
 [1 4 0 0 0]
 [0 4 0 0 0]
 [0 4 0 0 0]]

The accuracy of the model increases from 0.5 to 0.54 on using the scaled data. Neural Networks use gradient-based optimisation so scaled data makes the model more stable.

4k)

We trained three classifiers

Parameters were chosen based on Grid Search with a cross-validation of 5.

Surprisingly, on trying scaled and normalised data, the results were worse(0.65 accuracy) than using regular data. Using the Deviation Mean improved them a bit more(0.76 accuracy) but the regular data still had the highest accuracy.

The intermediate classifiers were judged using the accuracy, precision and the confusion metrics.

4l)

Bigger data sets with more samples for all the classes could help us build a better model.

QUESTION: 5

Clustering

5.a)

id and num are not useful for clustering.

id is just a unique identifier for each row, so it doesn't contain any useful information for clustering. num is the target column, since it is the column we are trying to predict, so it doesn't contain any useful information for clustering.

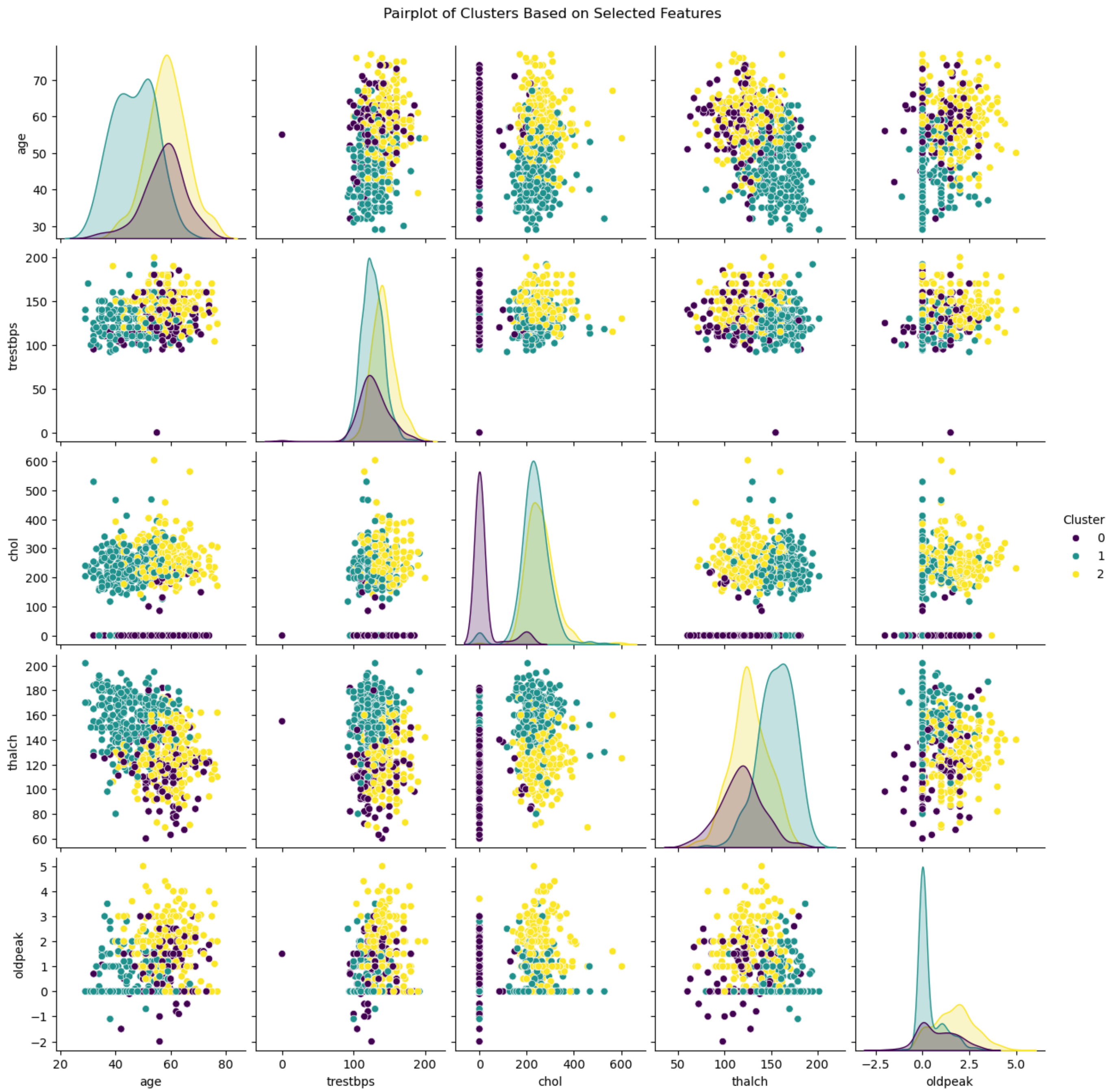
5.b)

```
Cohort Sizes:
Cluster
1    296
2    259
0    135
Name: count, dtype: int64

Cohort Centroids:
         age   trestbps        chol      thalch   oldpeak
0  57.548148  128.207556   19.659259  117.988593  0.816252
1  46.755102  126.485714  233.616463  155.628707  0.346779
2  58.551724  141.905556  256.924138  126.887663  1.694410
```

5.c)

As stated in the assignment, the feature chol has a particular influence on the clustering!

The clustering is not functioning correctly because of the extremely high density of zero values in chol. As a result, the results are skewed because the 0 value in clustering is overemphasized. We can use the **imputation method** (for instance, replacing the 0 value with the column mean value) **or eliminate the rows in chol that have a 0 value** to lessen this influence.



5.d)

A proper supervised learning algorithm would consider the relationship between the features and num directly, leading to more accurate predictions. Therefore, methods like decision trees, random forests, or logistic regression are more suitable for this task.

```
Confusion Matrix:
 [[11  5  0  0  0]
 [ 5 12  0  0  0]
 [ 1  4  0  0  0]
 [ 0  4  0  0  0]
 [ 0  4  0  0  0]]
Accuracy: 0.5
```